2016

Three Decades of Reliability in Communication Content Analyses: Reporting of Reliability Statistics and Coefficient Levels in Three Top Journals

Jennette Lovejoy
University of Portland, lovejoy@up.edu

Brendan R. Watson

Stephen Lacy

Daniel Riffe

Three Decades of Reliability in Communication Content Analyses: Reporting of Reliability Statistics and Coefficient Levels in Three Top Journals

Many scholars, editors, and manuscript reviewers would argue that failure to assess coding reliability is a "fatal flaw" in content analysis research. Establishing reliability helps readers evaluate the validity of data, as high reliability is a necessary, albeit insufficient, condition for validity, and provides a better basis for replicating a study (Riffe, Lacy, & Fico, 2014). There is evidence that reliability reporting may be improving. Lombard, Snyder-Duch, and Bracken (2002) and Snyder-Duch, Bracken, and Lombard (2001) found the percentage of studies assessing reliability was 69% for 200 content analysis articles indexed in *Communication Abstracts* during the 1994-1998 period. Riffe et al. (2014) reported that 74% of 80 *Journalism & Mass Communication (JMCQ)* content analyses published between 1998 and 2004 provided reliability assessments. More recently, Lovejoy, Watson, Lacy, and Riffe (2014) found that 76% of 581 content analyses in *JMCQ, Journal of Communication,* and *Communication Monographs* reported coding reliability assessment of some kind.

However, even if these data points reflect a trend toward increased reliability reporting, the adequacy of the reporting may be problematic. To describe a study's overall reliability, for example, one should report reliability statistics for each variable. However, Riffe et al.'s (2014) analysis of *JMCQ* found only 54% of studies provided test results for all variables "or at least provided the range of results for the relevant variables" (p. 122). Though their focus was primarily on reliability sampling procedures, Lovejoy et al. (2014) noted that only 27% of content analyses in their study included reliabilities for all study variables (only a third explicitly described the reliability

sample). In addition, reports should include reliability coefficients that take chance agreement into consideration (Krippendorff, 2013; Riffe et al., 2014). Yet Riffe et al. (2014) found fewer than half (46%) of 1998-2004 *JMCQ* content analyses reported coefficients that correct for chance. Lovejoy et al. (2014) also found chance-corrected reliability coefficients in fewer than half (49%) the content analyses they examined.

Although multiple studies have analyzed reporting of intercoder reliability in published content analyses (Feng, 2014; Lombard et al., 2002; Lovejoy et al., 2014; Pasadeos, Huhman, Standley, & Wilson, 1995; Riffe et al., 2014; Riffe & Freitag, 1997; Snyder-Duch et al., 2001), most were purely descriptive, using small samples, limited time periods, and limited numbers of journals. In addition, with one exception (Lovejoy et al., 2014), no published study examined changes in reliability reporting of published content analyses over time, and Lovejoy et al. (2014) examined reliability *sampling* procedures but did not characterize trends in reporting reliability coefficients for study variables, as will be presented in this study. The literature lacks an expansive, longitudinal examination of how reliability assessment is reported in major communication journals.

This study examines the recent history of reliability coefficient reporting in the three flagship journals of the Association for Education in Journalism and Mass Communication (AEJMC), the International Communication Association (ICA), and the National Communication Association (NCA). These journals arguably represent the highest standards of research in communication. The study uses a census of two journals and a representative sample of the other from 1985 to 2014 to assess recent and current reliability statistic reporting (e.g., percentage of simple agreement and reliability

coefficients), acceptable level of agreement, and whether coefficients were reported for every variable in the study, as well as how reporting practices have changed over time. This investigation will illuminate trends in reliability reporting, identify reporting practices that may require remediation to enhance the scientific rigor of content analysis methodology, and may inform future conversations about what the field's standards *should* be.

Literature Review

Agreement and Reliability Coefficients

Content analysis intercoder reliability assessments summarize the extent to which two or more coders, applying the same measurement tool (coding protocol) classify content units into the same categories. Such classification should employ trained coders applying a well-tested coding protocol with rules "that bind the researchers in the way they define and measure the content of interest" (Riffe et al., 2014, p. 98). Because of the importance or replication in social science, reliability must be vested in the protocol, which makes it possible for coders other than the original set to create reliable and valid data from the same protocol (Lacy et al., 2015).

Assessments of reliability have ranged from percentages of observed agreement (e.g., a pair of coders made identical classifications in 90% of trials) or what is labeled "simple agreement," to reliability coefficients that take into account the possibility of what represents false agreement among coders. The concept of "false agreement" merits comment. A false agreement is an agreement reached by coders that did *not* result from the application of a protocol by trained coders. Consider a content unit and a variable with categories A, B, and C. Although the protocol may clearly indicate that the unit

should be classified as an "A," it is possible for two coders to misapprehend or misapply the protocol and *both* mis-classify it as a "C." Their "agreement" is invalid according to the protocol, and thus, false. It is an agreement that another set of coders probably would not reach because it resulted from something other than applying the protocol consistently. Zhao (2012) calls this "erroneous agreement" but does not emphasize its connection to the protocol.

Holsti (1969) notes that simple agreement can overestimate reliability because as the number of coding categories for a variable declines, the likelihood increases that assignment of a content unit to the same category by coders could occur by "chance." The use of the term "chance" is unfortunate because such false agreements are not necessarily the results of random behavior (Gwet, 2008; Riffe et al., 2005). Although false agreements could be due to chance decisions by coders when applying the protocol, coders rarely randomly assign content to categories within variables if protocols contain adequate coding instructions and sufficient training has occurred. False agreements are more likely to result from poor or limited training, the lack of a common frame of reference, language skill differences among coders, etc. (Riffe et al. 2014).

Moreover, such false agreements are more serious that disagreement errors because *false agreements cannot be detected or differentiated from protocol-based agreements* and because false agreements have a high probability of creating data with validity problems. The fact that they cannot be detected is the reason scholars have argued that reliability is not the same as agreement (Cohen, 1960; Gwet, 2014; Krippendorff, 2013; Riffe et. al., 2014; Scott, 1955) and why reliability measures must adjust for "expected agreement" or "chance."

Though the present study and Feng's (2014) data show continued use of simple agreement, most scholars agree that *relying on simple agreement alone* is inadequate for assessing reliability (Cohen, 1960; Gwet, 2014; Krippendorff, 2013; Riffe et. al., 2014; Scott, 1955). Zhao (2012) does argues that simple agreement is a better measure of reliability than reliability coefficients that use probability for adjusting for false agreements. Some scholars have called for reporting simple agreement along with a reliability coefficient (Lacy et al., 2015; Riffe et al., 2014) as a way of providing more information to evaluate whether reliability is sufficient and to assist with replication.

Regardless of one's position on reporting simple agreement, debate persists on which reliability coefficients are optimal (Gwet, 2008; Krippendorff, 2012; Krippendorff, 2004a; Lombard, Snyder-Duch, & Bracken, 2004; Zhao, 2012; Zhao, Liu, & Deng, 2013). Existing reliability coefficients aim to estimate false agreement using either the distribution of agreements and/or disagreements to calculate "expected agreement," which is supposed to adjust for false agreement. It is the variations in how the coefficients estimate this expected agreement, and the assumptions that go with these variations, that contribute to the debate as to which reliability coefficient is superior. Although as many as 18 "chance adjusted" coefficients (Zhao et al., 2013) have been used in content analysis studies, data presented below indicate that communication studies tend to report one of three coefficients— Scott's Pi (1955), Cohen's Kappa (1960), and Krippendorff's Alpha (1980). Some scholars (Gwet, 2014; Zhao, 2012, Zhao et al., 2013) suggest these three are basically variations on the same formula.

However, Krippendorff (2013) argues that Alpha is a more useful coefficient than others because it adjusts for small sample sizes and can be used with more than two

coders and all levels of measurement (nominal, ordinal, interval, and ratio). He criticizes Kappa as being insensitive to small sample sizes and for how its expected disagreements is calculated and notes that Kappa will be higher than Alpha when the marginal sums of the contingency tables are unequal. He summarized the difference: "K overestimates reliability when coders have unequal proclivities for coding categories" (Krippendorff, 2014, p. 304). However, when the coders' "proclivities" are equal, Kappa and Alpha are roughly the same. Zhao, Liu, and Deng (2013) argued that Alpha favors small samples and, compared to other coefficients, creates smaller reliability coefficients, all else being equal, as sample sizes increases. Krippendorff (2013b) responded that Alpha corrects for biases that exist in coefficients such as Pi when samples are smaller.

Another controversy involves Alpha, Pi, and Kappa and the fact that they can produce very low coefficients even when levels of simple agreement are high (Feng, 2014; Gwet, 2008; Zhao et al., 2012), which can occur when data distributions are skewed (e.g., most of the coded units are in one category; see Riffe et al., 2014, pp. 119-120). Krippendorff (2013a) labels this "insufficient variation" (p. 319), writing that such data ". . . cannot be correlated with anything either, their analytical meanings are largely void, and they cannot convey sufficient information from the analyzed text to the research question" (p. 320).

That conclusion seems to ignore the fact that there have been, and will continue to be, populations with skewed distributions of categories that are nonetheless important to study. For example, Robinson and Anderson (2006) studied portrayal of older characters in animated children's television. Only 8% of characters were older and of these 107 characters, only1% were African American. The authors reported simple agreement to

assess reliability. Monk-Turner, Heiserman, Johnson, Cotton, and Jackson (2010) found only 5% of primetime TV characters were Hispanic and fewer than 2% were Asian American. As with Robinson and Anderson, the article reported only simple agreement. The authors do not report why they did not provide chance-corrected reliability coefficients, but it may be because of the skewed distribution phenomenon.

Addressing the phenomenon, Potter and Levine-Donnerstein (1999) argued that expected agreement should be calculated by using the normal approximation to the binomial distribution (rather than using the formulas for Pi, Alpha, and Kappa). However, this approach does not address any role of the protocol in disagreements and false agreements.

More recently, Gwet (2008, 2014) developed coefficient AC_1 for inter-rater agreement in health diagnoses. He reasoned that not all decisions are of equal difficulty, and AC_1 divides decisions into hard-to-score and easy-to-score. Gwet (2014) found that AC_1 results in coefficients that are lower than simple agreement but higher than Alpha, Kappa, and Pi. Krippendorff (2013a) has criticized AC_1 as difficult to interpret, a charge echoed by Ejima, Aihara, and Nishiura (2013).

If one accepts the need to adjust for false agreements, it appears that among the most commonly used coefficients Alpha has advantages over Pi and Kappa but that AC_1 can provide a possible substitute when skewed data distributions results in simple agreements that are significantly larger than the reliability coefficients. However, determining whether AC_1 is an acceptable general replacement for Alpha, Kappa, and Pi is beyond the scope of this study. Gwet (2008, 2014) has run Monte Carlo studies comparing coefficients, and Zhao (2012) ran an experiment using lengths of lines to test

coder agreement. However, these studies have not included the use of content protocols to guide coding of symbolic content. The call for further empirical study (Zhao, 2012) should be taken up by scholars. Until such studies help resolve the debate, we agree with Riffe, et al. (2014) and Lacy, et al. (2015) that manuscripts report Alpha, simple agreement, and AC_1 when data are skewed. Because the argument against Alpha, Pi and Kappa is that they tend to underestimate "true" reliability (Gwet, 2014; Zhao, 2013; Zhao et. al., 2013), using a more "strict" measure, such as alpha, helps to guard against confirmation bias in content analysis.

Number of Variables

Even researchers who report reliability occasionally do so in problematic ways. Feng (2014) examined articles in four journals between 1980 and 2011, including *JMCQ* and *Journal of Communication*: "(S)ome presented reliability for each variable while some just gave one single value by collapsing variables into one. Some of them did not report an exact value, but a range" (p. 8). In a study of 1998-2004 content analyses in *JMCQ,* Riffe et al. (2014) found "only 54% of the studies provided the test results for all variables or at least provided the range of results for the relevant variables" (p. 122).

Reporting an "average" or range of coefficients is problematic because variables with low levels of reliability can be "masked" (Riffe et al., 2014). Consider: A reported average of .85 among three Scott's Pi coefficients of .85+.85+.85 would generally be acceptable, but an average based on Pi coefficients of .95+.95+.65 includes one problem variable. Hiding variables with weak reliability prevents reviewers and readers from properly evaluating the reliability and, therefore, validity of individual variables, as reliability is a necessary precursor of validity.

More practically, replication in research requires that researchers be able to evaluate *each* of the measures used originally. If the same measures are used, reliability coefficients from the original study can help clarify variation in results between the original and the replication. Just as publication of questionnaire data that measure latent variables requires reporting of each scale's reliability (e.g., Cronbach's alpha), a similar standard should apply for content analysis variables.

Acceptable Reliability Levels

But if one chooses Scott's Pi, Krippendorff's Alpha, Cohen's Kappa, or another content analysis reliability coefficient, how high is "high enough"? This is not an easy question because different types of content variables fall on a continuum from easy to difficult. But coding difficulty is a function of a variety of factors (Riffe et al., 2014). These include: the nature of the variable (e.g. valence toward a person or classifying content units into categories such as topic), the quality of the content (e.g., adherence to grammar and syntax standards), the quality of the protocol (e.g., adequate guidance for coders), and the quality of coder training (e.g., coders' ability to apply the protocol consistently).

Riffe et al. (2014) divide content units into physical and meaning. The former involve "mechanical" measurement of discrete units such as minutes, square inches and numbers of television programs. "Meaning" content units carry semantic meanings, such as types of sources quoted in a news article or tone toward a candidate. Measuring physical units is straightforward and should yield high levels of reliability (e.g., higher than .90). Meaning units, some of which involve counting, can vary in difficulty to a greater degree than physical units.

However, the fact that coding differs from one variable to the next, does not automatically call for lower minimum levels of reliability for some variables. First, the difficulty of coding certain types of variables can be compensated for by improving the protocol or better coder training. Second, the generally accepted levels (discussed below) are in fact not particularly high. As Krippendorff (2004b) said: "Even a cutoff point of α = .80—meaning only 80% of the data are coded or transcribed to a degree better than chance—is a pretty low standard by comparison to standards used in engineering, architecture, and medical research" (p. 242).

But what should be the minimum acceptable levels for reliability coefficients? Some scholars offer tentative guidelines. Wimmer and Dominick (2003) estimated that most published content analyses report coefficients of .75 or above when using Pi or Alpha (p. 159). Riffe, Lacy, and Fico (2005) recommended reliabilities in the .80-.90 range because most published content analyses they examined reported .80 or higher. Neuendorf (2002) described the Riffe, Lacy, & Fico (1998) .80-.90 prescription as "a relatively high standard" (p. 143).

Kaid and Wadsworth (1989) note that "researchers can usually be satisfied with coefficients over +.85, while those below +.80 should be suspect" (p. 209). Yet Krippendorff (1980) described a study in which he reported variables with Alphas above .80 *and* in which he used Alphas between .67 and .8 "for drawing highly tentative and cautious conclusions," a rule of thumb that "might serve as a guideline elsewhere" (p. 147).

Riffe et al. (2005) state that their .80-.90 range also is "appropriate for Scott's Pi with nominal data and a large sample" (p. 151). But Wimmer and Dominick (2003, p.

159) suggest .75 as the minimum for Pi and Alpha. Popping (1988) calls for minimum values of .80 for Kappa, though Bannerjee, Capozzoli, McSweeney, and Sinha (1999) propose a minimum Kappa of .75.

Having reviewed various coefficients, Neuendorf (2002) concluded that simple agreement levels of 90% or higher are absolutely acceptable and simple agreement levels of 80% or higher should be acceptable for most variables; chance-corrected statistics, such as Pi and Kappa, "are afforded a more liberal criterion" (p. 143).

Though these "rules of thumb" (Neuendorf, 2002, p. 143) may be useful and easily remembered, they clearly acknowledge and accept a certain amount of unreliability and, therefore, error, in coding—80% is not, obviously, 100%, and .8 is not 1.0. As with all scholarly rules of thumb, research needs to examine the relationship between reliability levels and the validity of conclusions from the data. At what reliability level does the conclusion drawn from the data create invalid conclusions?

Research Questions

Six research questions were proposed, exploring the frequency and type of reliability assessment in content analyses of communication content published in three leading communication research journals. For each **RQ**, variations across journals and time periods will be evaluated.

RQ1: What types of statistics (simple agreement or reliability coefficient) were used during the 1985-2014 period in the three flagship journals to represent reliability of content analysis?

RQ2: Were reliability coefficients reported for every variable?

RQ3: Which reliability coefficients were reported in content analysis articles?

RQ4: How many articles reported one or more reliability coefficients between .70 and .79?

RQ5: How many articles reported one or more reliability coefficients below .70?

RQ6: How well do year and journal predict the trends in reliability reporting?

Method

A content analysis was used to systematically examine representative samples of issues of three major communication research journals drawn from a 30-year period (1985-2014): *Communication Monographs* (*CM*), published on behalf of the National Communication Association; *Journal of Communication* (*JoC*), published for the International Communication Association; and *Journalism & Mass Communication Quarterly* (*JMCQ*), a journal of the Association for Education in Journalism and Mass Communication. All publish quarterly issues, although *Journal of Communication* expanded to six issues a year in 2011, and include multiple articles in each issue. As flagship journals for the largest three communication associations, these journals likely carry research of high quality, if not the highest quality. Limiting the study to these three also made the coding manageable.

Each sampled issue of the three journals was screened by two coders to determine which articles qualified as content analysis, using four criteria were:

1. At least some data analyzed for the article were obtained by examining existing content (mediated or interpersonal) or content created specifically in response to experimental stimuli. Other, non-content data, can be used in the article and it still is classified as a content analysis article (e.g., an agenda-setting study matching content data with survey data).

2. The content must be divided into discrete measurement units in order to assign numbers for quantitative analysis (i.e., a historical, legal, or qualitative study, or essay, based on a reading of all texts that include a key term, is not a content analysis).

3. The content data do not have to have been collected by author(s) for the article to count as a content analysis article (e.g., analyses of previously collected content data would qualify).

4. The content analysis must deal with the assignment of data based on meaning of the symbols. For example, a study that measures the square inches of "city" content on a website would qualify because the content being measured is defined by the meaning associated with the names of particular cities. Measurement of all advertising space or text space in a newspaper is independent of meaning and would not be a content analysis.

Initial examination of the articles thus identified revealed that *JMCQ* carried a considerably larger number of content analysis articles than the other journals. Thus, while all content analysis articles from *CM* (N = 153) and *JoC* (N = 193) in the 30 years were included, two issues per each of the 30 years were randomly selected from *JMCQ*, with the 60 issues yielding 326 *JMCQ* content analysis articles.

The reliability of the identification screening was tested. The 60 sampled issues of *JMCQ* included 985 total articles. Krippendorff's Alpha was .92 (simple agreement=97%) for inclusion of *JMCQ*'s 326 content analyses, .83 (simple agreement=95%) for *JoC*'s 193 content analyses (from 1,233 total articles), and .88 (simple agreement=96%) for *CM*'s 153 content analyses (from 708 total articles).

The 1985-2014 period represented years during which three major content analysis texts (Krippendorff, 1980, 2004b, 2013; Neuendorf, 2002; Riffe et. al., 1998; Riffe et al., 2005, 2014) were introduced and/or revised. All three encourage use of reliability coefficients that consider chance agreement. Krippendorff's first edition was published in 1980, but because texts are not universally adopted immediately, 1985 was selected as the starting point. Reliability in content analyses were examined through 2014 to provide data on the most current practices.

A coding protocol and variable definitions (available upon request) were refined through several rounds of training, practice sessions, and test coding by three of the study's authors on randomly selected articles not used in the study but drawn from the three journals. After protocol modifications, two coders (one author who helped develop the protocol and one author who did not) performed three pilot checks (n = 20, 17, and 10 articles, respectively), again using randomly selected articles from all three target journals not in the sample; simple agreement exceeded 82% for all variables in all three pilot tests, a level deemed sufficient to begin coding the sample.

Then, the same two coders each independently coded half the 672 sampled study articles, with articles stratified by journal and randomly assigned. To assess intercoder reliability of the protocol for the actual study sample, the two coders double-coded 114 articles randomly chosen (*JMCQ* = 50, *JoC* = 38, *CM* =26) on the basis of the Lacy and Riffe (1996) formula for reliability sample size (assuming 95% probability and a 90% agreement level in the population). Final reliability coefficients were judged to be acceptable with all variables above .85 chance-corrected coefficient (Kaid & Wadsworth, 1989; Neuendorf, 2002; Riffe, Lacy & Fico, 2005) and are reported below. To illustrate

the similarities discussed above between three commonly cited chance-corrected reliability coefficients, we present reliability using Alpha, Pi, and Kappa. In addition, we include simple agreement reliability figures.

Simple Agreement Reported: (Alpha = .860, Pi = .859, Kappa = .860, Simple Agreement = 93.0%). Was a simple agreement reliability coefficient reported in the article? Simple agreement is sometimes called "Holsti's formula" or "Holsti's reliability coefficient." If the term "reliability coefficient" was used without a specific label (Scott's Pi, Krippendorff's Alpha, etc.), it was coded as simple agreement.

Type of Reliability Measure: (Alpha = .886, Pi = .885, Kappa = .886, Simple Agreement = 92.1%). Were coefficients reported that correct for chance and which (Scott's Pi, Krippendorff's Alpha, Cohen's Kappa, Gwet's Gamma, Benini's Beta, Guttman's Rho, etc.) was reported for the "final" (non-training, non-pilot) reliability test? These coefficients had to be explicitly identified by name or by an explanation of how chance agreement and the coefficient were calculated.

Reliability Coefficients Reported for All Study Variables: (Alpha = .852, Pi = .851, Kappa = .851, Simple Agreement = 90.4%). Was a reliability coefficient reported for each variable (in the text, tables, endnotes, footnotes, or appendices)? Variables that involve transcription, such as title, page number, date, etc. did not have to have a reliability measure reported, but all variables that involve the meaning of symbols did.

Levels of Reported Reliability Coefficients: (Alpha = .917, Pi = .917, Kappa = .917, Simple Agreement = 93.9%). What was the lowest chance-corrected reliability coefficient reported? If only one reliability coefficient was reported, this was recorded as the lowest coefficient. For analysis, these values were re-coded into mutually exclusive

categories (e.g., lowest reported coefficient was <.70, was between .70 and .79, or was ≥.80, etc.).

Fewer than 6% of all articles reported Pearson product-moment correlations as reliability statistics. Krippendorff (2013) argues that tests of association, such as Pearson's *r,* are not equivalent to tests of agreement. Correlations can be fairly high and still have a high percentage of actual disagreements. For example, if two coders analyze 10 content items using a five-point scale, it is possible to have zero agreement but a 1.0 Pearson's correlation because a correlation deals with consistent patterns in data and not agreement. On the other hand, Riffe et al. (2014) argue that correlations can be appropriate with physical variables, such as number of words, sentences, or paragraphs, that are "mechanical" in their recording, and do not involve interpreting symbolic meaning. Because the appropriateness of using correlation was not always obvious in the study articles, it was treated as a reliability coefficient in some of the analysis, which admittedly represents a liberal interpretation of its use as a reliability coefficient.

Data Analysis

In order to analyze the trend across time and by journal, the percentage of articles that contained a given variable (type of reliability coefficient, etc.) was calculated for each year and journal. To answer all **RQ**s except **RQ3** and **RQ6**, the percentages per year within each journal were analyzed using OLS simple linear regressions of the dependent variable on year. The goodness of fit for the regression equation is reported using r-square. A significant positive r-square indicates that the annual percentage of articles increased during the 30-year period. A significant negative r-square indicates the percentage declined during the period, and a small insignificant correlation suggests little

or no change. **RQ3**, asking which coefficient (Alpha, Kappa, Pi, other chance-corrected coefficient, or correlation) a study reported, was not explored using simple regression because of early study years containing a small number of, or no, articles that reported a particular reliability coefficient.

RQ6 was answered using multivariate ordinary least squares regression. The dependent variables are the percentage of articles in a year containing the characteristics mentioned in the **RQs** above. Independent variables were publication year and two dummy variables, one for *JMCQ* and one for *JoC*, with *CM* used as the reference group and assigned a zero for both dummies. The impact of year and journal is reported using part (semi-partial) correlations, which when squared show the unique variance shared between the individual independent variable and the dependent variable. However, because scholars recommend against reporting *only* simple agreement in content analyses, no model was run using percentages of articles reporting such flawed assessment. Finally, no multivariate model examined percentages of Alpha, Kappa, Pi, other chance-corrected coefficients, and correlations for aforementioned reasons.

Visual examination of the data did not suggest curvilinear relationships between time and any dependent variable, though sample size precluded formal tests of these relationships (e.g., quadratic or cubic effects of publication year). To account for variability in the number of content analysis articles in a given journal per year, we conducted all regression analyses controlling for this variable and results were unchanged (data not reported). We thus report findings from univariate OLS regression for **RQs 1**, **2**, **4**, and **5** and multivariate OLS regression for **RQ6** that includes as covariates: a linear term for year and the aforementioned dummy variables for journal.

The data were tested for violations of regression assumptions. First, all variables had skewness of less than 1, except for percentage of articles with coefficients below .7, and percentage with one or more coefficients between .7 and .79. Both had one outlier case (defined as more than three standard deviations from the mean); the outliers were re-assigned the value of the largest case *below* three standard deviations from the mean. The skewness measure became 1.09 for coefficients between .7 and .79 and 1.37 for coefficients below .7. The data met all other tests.

Because years were used, Durbin-Watson tests were run for the regression equations. The Durbin-Watson coefficients for each regression, identified here by the dependent variable in each, were the percentages: 1. of articles with reliability coefficients – 1.29; 2. of articles with coefficients reported for each variable – 1.44; 3. of articles with one or more coefficients below .7 – 1.76; and 4. of articles with one or more coefficients between .7 and .79 – 1.83. The lower limit for a regression with n = 90 and three independent variables was 1.59 and the upper limit was 1.73 (Mansfield, 1987, p. A26). There are no negative autocorrelations, but we cannot reject the null hypotheses that variables 1 and 2 do not have slight positive autocorrelations.

These variations from normality for two variables and autocorrelation for two dependent variables are not major concerns. First, regression is robust for minor violations of assumptions (Chatterjee & Price, 1977, p. 9). Second, assumptions about both normality and autocorrelation concern biased estimation of parameters (Bowerman, et al., 1986, pp. 551-562). The coefficient estimates for the sample are not affected. Given that the sample was taken from 84% of the total journal issues during this period

(308/368), the part correlations from the regression equation will be interpreted as suggestive for further study.

Results

RQ1 asks what types of statistics (simple agreement or reliability coefficient) were used during the 1985-2014 period to report content analysis reliability. Figure 1 and Table 1 data provide part of the answer, examining trends in reporting simple agreement in the three journals. *JMCQ* data showed an apparent overall decline, from 42% in the 1985-1989 period to 19% in 2010-2014. R-square for year and percentage of articles reporting simple agreement was significant (r-square = .17, p = .024). Similar results were found with *JoC*, though trends in reduced reporting of simple agreement did not reach statistical significance for *JoC* due to large variation across years (r-square = .11, p = .073). Fifty-three percent of content analyses published in *JoC* from 1985-1989 reported simple agreement, while 26% reported simple agreement between 2010 and 2014. The relationship between year and percentage of articles reporting simple agreement for *CM* was also negative, and statistically significant (r-square = .19, p = .015). Overall, the reporting of simple agreement declined during the 30-year period, but more strongly in *JMCQ* and *CM* than *JoC*. Of course, declining levels of reporting simple agreement are not problematic, *if* chance-corrected coefficients are being used instead of or with simple agreement.

INSERT TABLE 1 AND FIGURE 1 ABOUT HERE

In fact, Figure 2 and data in column two and three of Table 1 examine reporting of reliability coefficients during the 30-year period, further addressing **RQ1**. *JMCQ* showed an overall increase in the percentages of articles reporting a reliability

coefficient, from 10% in years 1985-1989 to 75% in years 2010-2014. The r-square for the 30-year period was positive and significant (r-square = .72, $p < .001$). *JoC* had a similar trend, from 16% in years 1985-1989 to 78% in years 2010-2014, with five of the last eight years studied reaching 100% (r-square = .61, $p < .001$). However, *CM* showed no clear-cut trend (r-square = .03, $p = .386$): while 1985-1989 had 65%, 68% in years 2010-2014 reported a reliability coefficient, with middle 5-year time periods ranging from 67% to 79%. Overall, the percentage of articles reporting reliability coefficients increased for *JMCQ* and *JoC*, but *CM* showed no statistically significant pattern. *JMCQ* and *JoC* trends notwithstanding, one could argue that anything less than 100% is inadequate.

By comparing column two and three, it becomes evident that most of the correlations used for reliability were found in *CM*. The percentage of *CM* articles during the most recent period (2010-2014) with coefficients when correlations were included equaled 68%, but when the correlations were dropped, the percentage with coefficients equaled only 57%. This pattern was clear throughout the 30 years. Few of the *JMCQ* and *JoC* articles used correlations. The researchers did not anticipate such a high concentration of correlations in one journal and did not code for whether the variables using correlations were physical or meaning units. This finding warrants additional study.

INSERT FIGURE 2 ABOUT HERE

RQ2 asks if reliability coefficients were reported for each variable. Only 24.9% of all study articles reported a reliability coefficient for each variable across the 30-years, but Figure 3 and Table 1 illustrate how the percentages for each journal changed over time. For *JMCQ*, only 2% of articles between 1985 and 1989 reported a reliability

coefficient for every variable, and the proportion improved to only 6% by the 1995-1999 study period. By the final 5-year time period, however, the percentage reached 50%. The association between publication year and reporting a reliability coefficient for every variable was positive and significant (r –square = .58, $p < .001$). An even stronger trend appeared for *JoC* (r-square = .64, $p < .001$). From 1985-1989, the percentage of articles reporting a reliability coefficient for each variable was 3%. By the 2005-2009 study period, the proportion of *JoC* content analysis articles reporting reliability for every variable reached 64%, with only slightly lower reporting at 59% from 2010-2014. The pattern toward increased reporting was not as strong for *CM*. The proportion of articles reporting reliability for all variables was 27% from 1985 to 1989, then peaked at 46% from 2005-2009 before dipping to 32% in the final 5-year study period. This pattern represented a non-statistically significant change over time (r-square = .07, $p = .153$). Overall, the 30-year period showed an increase in the annual percentage of articles reporting a reliability coefficient for each variable, but with smaller percentages for *CM* than *JMCQ* or *JoC*.

INSERT FIGURE 3 ABOUT HERE

RQ3 asks which specific reliability coefficients were reported. Across journals, the most often used coefficient was Scott's Pi (15%), followed by Cohen's Kappa (14%) and Krippendorff's Alpha (5%); 4% of articles included another chance-corrected reliability coefficient (e.g., Gwet's Gamma, Benini's Beta, Guttman's Rho) and the majority (57%) reported no reliability coefficient. Table 2 shows how the reporting of these five coefficients varied across time. Pi, Kappa, and Alpha were used consistently and generally grew in use over the study period, with 77% of articles from 2005-2009

and 73% of articles from 2010-2014 reporting a chance-corrected reliability coefficient, most commonly Pi, Kappa, or Alpha. An inverse trend was observed for lack of reliability reporting, with 81% of reviewed articles from 1985-1989 failing to report a reliability coefficient, reducing to 17% of articles form 2005-2009 and 23% of articles from 2010-2014. Comparatively speaking, Krippendorff's Alpha was used less often until the most recent 5-year study period (2010-2014), at which point its use surpassed that of Scott's Pi and Cohen's Kappa. Variations in use among the three coefficients may reflect how long they have been available. Pi was introduced in 1955 (Scott, 1955), and Kappa five years later (Cohen 1960). Alpha appeared two decades later (Krippendorff, 1980). In addition, Alpha is more difficult to calculate by hand, which could have delayed its growth until recent years. An SPSS macro (Hayes & Krippendorff, 2007b) and online calculator (Freelon, 2013) now make calculation easier.

INSERT TABLE 2 ABOUT HERE

Because scholars (Kaid & Wadsworth, 1989; Popper, 1988; Riffe et al., 2005) suggest that reliability coefficients below .80 can be problematic, **RQ4** asked how many sampled articles reported one or more reliability coefficients between .70 and .79. Figure 4 and Table 1 show variable reporting of coefficients in this range. For *JMCQ,* reporting of reliability between .70 and .79 was uncommon for the first 20 years of the study period (1985-2004), with 5-year percentages ranging from 3% to 8%. From 2005-2009, the percentage had increased to 21% but declined to 6% in the final study period. The time trend in reporting reliability between .70 and .79 for *JMCQ* was not statistically significant (r-square = .05, p = .221). A similar and also non-significant pattern (r-square = .10, p = .082) was seen with *JoC*. The proportion of articles reporting coefficients in

the .70-.79 range in the first 20 years was between 0% (1990-1994) and 6% (1985-1989). By the 2004-2009 period, the percentage had increased to 24% but declined to 14% in the final study period. *CM* also showed no relationship between year and percentage of articles with coefficients in this range (r-square = .03, p = .368), though reporting of coefficients between .70 and .79 was more common in *CM* than in *JMCQ* or *JoC*. For the 30-year period, 23% of *CM* articles had coefficients within this range, compared to 10% for *JoC* and 6% for *JMCQ*.

In summary, the three journals contained a modest percentage of articles with problematic reliability coefficients between .7 and .79, although with considerable variance. These reporting practices across all journals appeared to peak from 2005-2009. No journal, however, exhibited growth in in the percentage of such articles over time.

INSERT FIGURE 4 ABOUT HERE

RQ5 goes even further in examining problematic reporting, asking how many articles reported one or more reliability coefficients *below* .70. Of the 326 *JMCQ* articles, 4% reported coefficients of less than .70. Data in Figure 5 and Table 1 show an increase in such articles during the 30-year study period, although this time trend did not reach statistical significance (r-square = .12, p = .057). Of the 193 *JoC* content analysis articles, 12% had one or more reliability coefficients below .70. There was a statistically significant increase in reporting of reliability coefficients below .70 in *JoC* over the study period (r-square = .28, p = .002). Of the 153 *CM* articles, 12% contained one or more coefficients below .70. For *CM*, there was a statistically significant negative trend in reporting of coefficients below .70 (r-square = .16, p = .028). Only 4% of *CM* articles reported reliability coefficients below .70 in the last 10 years of the study. Overall, *JoC*

showed increases in the reporting of variables with questionable reliability, while *JMCQ* showed a trend in this direction. Even though *CM* showed a modest decline, even its 12% of articles reporting sub-.70 coefficients seems excessive.

INSERT FIGURE 5 ABOUT HERE

RQ6 asks how well year and journal predicted trends in reliability reporting. The semi-partial correlations and regression coefficients in Table 3 indicate variable predictive utility of year and journal for the four dependent variables. The best predictor of the annual percentage of articles reporting a reliability coefficient and the annual percentage of articles reporting a reliability coefficient for all study variables was publication year (r-square = 0.469, $p < 0.001$ and r-square = .427, $p < .001$, respectively). Publication year was associated with increased reporting of reliability coefficients for at least one and for all study variables, after controlling for journal. Publication year, however, was unrelated to the percentage of articles reporting coefficients below .70 or in the .70-.79 range (semi-partial = .130, $p = .221$ and semi-partial = .072, $p = .470$, respectively).

INSERT TABLE 3 ABOUT HERE

Journal was also related to the percentage of articles reporting reliability coefficients and the percentage of articles reporting coefficients for all study variables. After controlling for publication year, *JMCQ*, relative to *CM*, published a lower percentage of: 1. articles that reported reliability coefficients (semi-partial = -.295, $p < 0.001$); 2. articles that reported coefficients for each study variable (semi-partial = -.192, p - .021); and 3. articles with coefficients ranging from .70-.79 (semi-partial = -.354, $p = .001$). Similar patterns, but slightly smaller associations, were found with *JoC*. Relative

to *CM* and after controlling for publication year, *JoC* published a smaller percentage of articles that: 1. reported reliability coefficients (semi-partial = -.253, $p = .002$); and 2. reported reliability coefficients between .70 and .79 (semi-partial = -.298, $p = .004$).

Discussion

This content analysis of content analyses investigated reporting of coding reliability coefficients in the flagship journals of the three largest communication associations (AEJMC, NCA, and ICA). Data suggest improvements in reporting across time, but also identified areas for additional improvement. On the positive side, the percentage of articles in the journals that reported reliability coefficients suggests that reporting chance-corrected reliability coefficients should continue to improve with time.

However, the rate of increase varied across the study period, and by the 2010-2014 period, 23% of the *JMCQ* articles, 25% of the *JoC* articles, and 32% of the *CM* articles did not include a chance-corrected reliability coefficient, though content analysis reference works published during this study's 30-year timeframe suggest that *every* content analysis reliability check should use chance-corrected coefficients (Krippendorff, 2004b, 2013; Lacy & Riffe, 1993; Neuendorf, 2002; Riffe et al., 1998, 2005, 2014).

The relatively extensive use of Pearson's product-moment correlation within *CM* articles raises some concerns. Krippendorff (2013) rejects the use of correlations. Riffe, et al. (2014) argue it is acceptable when physical units such as minutes and square inches are coded. Exploring the type of variable evaluated with correlations was beyond the scope of these data.

In addition, the percentage of articles not reporting a reliability coefficient for *each* variable was worrisome, even though the percentage doing so increased during the

30 years. During 2010-2014, only 50% of *JMCQ* articles, 59% of *JoC* articles, and 32% of *CM* articles reported reliability coefficients for each variable. Failure to report a reliability figure for each variable may allow findings that may be tenuous—due to low reliability on a given variable—to *appear* stronger than they are. In addition, future researchers may erroneously replicate variables that did not reach acceptable levels of reliability but were "masked" by an average or range of coefficients. Reliability reporting for each variable is essential in identifying valid analyses and warrants table or endnote space, at minimum.

While the reporting of chance-corrected reliability coefficients increased during this 30-year time period, the use of Pi, Kappa, and Alpha varied across time. The most often used coefficient during the 2010-2014 period was Krippendorff's Alpha (28.8% of the articles) and the second was Cohen's Kappa with 25.2%. However, use of Pi declined during the 2005-2014 period from 28.9% to 16.2%. The growth of Alpha may reflect its flexibility and the fact that an SPSS macro was published to make calculation easier.

Although the presentation of data differed between this study and Feng's (2014), the descriptive results of the studies are generally consistent. However, while Feng referred to trends across time, he did not provide statistical analyses of those trends, nor did he provide statistical analyses of variation among journals.

As reporting of reliability coefficients increased, the reporting of simple agreement declined. However, during the 30 years, 36% of the *JMCQ* articles, 28% of the *JoC* articles, and 18% of the *CM* articles used only simple agreement to assess reliability. It is likely that some of these articles, as may be the case with two studies cited above (Monk-Turner, et al., 2010; Robinson & Anderson, 2006), reported simple agreement

because they involved the skewed distribution phenomenon. Exploring this relationship could yield an explanation, but not all articles report enough information to evaluate them.

Regression analyses showed a fairly strong positive relationship between year and reporting reliability coefficients and between year and reporting coefficients for each variable: As time passed, reporting became better. However, the increase was not consistent among journals. With *CM* as the reference group, *JMCQ* and *JoC* were slightly less likely to report reliability coefficients and one for each variable when controlling for time. Because the regression equations for these two content variables accounted for less than 50% of the variance, there are other variables that would play a role in predicting the dependent variables.

Of particular interest is the reporting of reliability coefficients with low reliability. Although there is no consensus about what level of reliability is acceptable, most texts suggest that coefficients greater than .8 are acceptable, that coefficients between .7 and .79 might be acceptable with reservations, and that coefficients below .7 should be used with extreme caution. On the basis of those guidelines, data from this study raise some concern. The percentage of articles reporting at least one chance-corrected coefficient smaller than .7 increased until 2010-2014, during which the percentage dropped considerably for *JMCQ* and *JoC*, although 12% of the *JoC* articles during this period had variables below .7. *JMCQ* had a slight negative relationship. In other words, *JMCQ* was less likely to publish articles with coefficients this low. This study could not detect the factors influencing the decline, but second editions of the Krippendorff (2004b) and Riffe

et al., (2005) text were published before this period and both caution against use of variables with reliability below .7.

There also was an increase in the percentage of articles per year reporting coefficients between .7 and .79. The regression equation using year and two journal dummy variables accounted for 15% of variance in reporting coefficients between .7 and .79. Time was not an important variable for predicting this percentage, but the percentage declined for *JMCQ* and *JoC* compared to *CM* even after controlling for year. The percentage for CM remained in double digits during the period until the final five years (2010-2014). It varied from a high of 42% in 1990-1994 to a low of 7% in 2010-2014.

Overall, the reporting of reliability coefficients improved during the 30-year period. However, further improvement is warranted. Every content analysis article should include chance-corrected reliability coefficients for every variable included in the analysis. The use of low-level reliability coefficients needs further study to determine if these lower levels of reliability represented truly exploratory research or a lack of diligence on the part of reviewers.

Of course, data from the present analysis cannot explain the changes found here, or how authors used particular variables in data analyses. However, the four regression equations showed that reporting a reliability coefficient and reporting a coefficient for each variable was predicted better by year than by journal, but predicting the presence of low reliability variables was predicted better by journal than by year. These results and the r-square levels suggest a study examining the causes behind reporting levels would be useful in improving reliability reporting.

The variations in use of reliability coefficients found across journals suggest two possible explanations, nature of content and training, although neither can be tested with these data. The variations might represent different levels of difficulty in coding audience-oriented communication versus interpersonal communication. Content created to serve the public usually follows routines for information gathering and models for presenting that information. For example, news writers apply a limited number of writing styles and structures, which provides a uniformity of content not found in written responses to experimental treatments. JMCQ primarily deals with public media more than JOC, which runs more articles about public media than does CR. These patterns reflect the research interests of members of the three associations.

A corollary of this argument is that variations in research training between communication doctoral programs and media/information programs could explain some of the variation among the journals. Scholarly articles and texts used in content analysis classes would typically be taken from the journals from the faculty members' associations. As a result, the use of a particular coefficient is encouraged from one generation of scholar to another. The time period in which a scholar trained also might play a role. Older scholars might be less likely to use Alpha than scholars who graduated during the past five or ten years.

In addition, the parameters for the time period suggest a possibility. The years were selected because of the publication of three content analysis texts. The adoption of any or all of these texts may have contributed to standardizing procedures in reporting reliability

Another interesting result was the growth in the use of Krippendorff 's Alpha. Its use increased considerably from only 3.4% in 1985-1989 to 28.8% in 2010-2014. This may reflect the growing use of Krippendorff's text, the argument by Hayes and Krippendorff (2007a) for its adoption as the omnibus coefficient, and the availability of an SPSS macro (Hayes & Krippendorff, 2007b) and online calculator (Freelon, 2013) to calculate Alpha.

As with all studies, this one has limitations. First, it involves only three journals. One would expect that the flagship journals for the largest communication associations would represent the best research, but this is itself an empirical question. If it is true that these journals publish high quality research, what are the reporting practices of the proliferation of specialized communication journals? How do these practices relate to journals' reputation and impact? Second, the sample included only communication journals. Other social sciences might vary in how they report reliability. This study assumes standards of reporting (both procedures and levels) that may not be universally accepted, an assumption that also suggests future study. A survey of scholars and teachers of content analysis about appropriate standards would help explain these results.

References

Bannerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond Kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*, 3-23.

Bowerman, B. L., O'Connell, R. T., & Dickey, D. A. (1986). *Linear statistics analysis: An applied approach*. Boston, MA: Duxbury Press.

Chatterjee, S., & Price, B. (1976). *Regression analysis by example*. New York: John Wiley & Sons.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Ejima, K., Aihara, K., & Nishiura, H. (2013). On the use of chance-adjusted agreement statistic to measure the assortative transmission of infectious diseases. *Computational and Applied Mathematics*, *32*(2), 303-313.

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *European Journal of Research Methods for the Behavioral and Social Sciences*, *11*(1), 13-22.

Freelon, D. (2013). ReCal: Reliability calculation for the masses [software]. Retrieved from http://dfreelon.org/utils/recalfront/

Gwet, K. L. (2014). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61,* 29-48.

Hayes, A. F., & Krippendorff, K. (2007a). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77-89.

Hayes, A. F., & Krippendorff, K. (2007b). KALPHA [software; SPSS macro]. Retrieved from http://www.afhayes.com/public/kalpha.zip

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley Publishing.

Kaid L. L., & Wadsworth, A. J. (1989). Content analysis. In P. E. Barker & L. L. Barker (Eds.), *Measurement of Communication Behavior*. New York: Longman.

Krippendorff, K. (2013a). *Content analysis: An introduction to its methodology,* 3rd ed. Los Angeles, CA: Sage.

Krippendorff, K. (2013b). Commentary: A dissenting view on so-called paradoxes of reliability coefficients. In C. T. Salmon (Ed.), *Communication yearbook 36* (pp. 481-499). New York: Routledge.

Krippendorff, K. (2004a). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30,* 411-433.

Krippendorff, K. (2004b). *Content analysis: An introduction to its methodology,* 2nd ed. Thousand Oaks, CA: Sage.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage.

Lacy, S., & Riffe, D. (1996). Sampling error and selecting intercoder reliability samples for nominal content categories. *Journalism & Mass Communication Quarterly*, *73*, 963-973.

Lacy, S., & Riffe, D. (1993). Sins of omission and commission in mass communication quantitative research. *Journalism Quarterly*, *70*, 126-132.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*, 587-604.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2004). A call for standardization in content analysis reliability. *Human Communication Research, 30,* 434-437.

Lovejoy, J., Watson, B. R., Lacy, S., & Riffe, D. (2014). Assessing the Reporting of Reliability in Published Content Analyses: 1985-2010. *Communication Methods and Measures*, *8*, 207-221.

Mansfield, E. (1987). Statistics for business and economics: Methods and approaches, 3rd ed. New York: W.W. Norton.

Monk-Turner, E., Heiserman, M., Johnson, C., Cotton, V., & Jackson, M. (2010). The portrayal of racial minorities on prime time television: A replication of the Mastro and Greenberg study a decade later. *Studies in Popular Culture*, *32*(2), 101-114.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage Publications.

Pasadeos, Y., Huhman, B., Standley, T., & Wilson, G. (1995). Applications of content analysis in news research: A critical examination. Paper presented at the annual conference of the Association or Education in Journalism and Mass Communication, Washington, D.C.

Popping, R. (1988). On agreement indices for nominal data. In W.E. Saris & I. N. Gallhofer (Eds.), *Sociometric Research: Data Collection and Scaling* (pp. 90-105). New York: St. Martin's Press.

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27,* 258-284.

Riffe, D., & Freitag, A. (1997). A content analysis of content analyses: 25 ears of Journalism Quarterly. *Journalism & Mass Communication Quarterly*, *74*, 873-882.

Riffe, D., Lacy, S., & Fico, F. G. (2014). *Analyzing media messages: Using quantitative content analysis in research* (3rd ed.). New York: Routledge.

Riffe, D., Lacy, S., & Fico, F. G. (2005). *Analyzing media messages: Using quantitative content analysis in research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Riffe, D., Lacy, S., & Fico, F. G. (1998). *Analyzing media messages: Using quantitative content analysis in research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Robinson, T., & Anderson, C. (2006). Older characters in children's animated television programs: A content analysis of their portrayal. *Journal of Broadcasting & Electronic Media*, *50*(2), 287-304.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321-325.

Snyder-Duch, J., Bracken, C. C., & Lombard, M. (2001). Content analysis in communication: Assessment and reporting of intercoder reliability. Paper presented at the annual conference of the International Communication Association, Washington, D.C.

Wimmer, R. D., & Dominick, J. R. (2003). *Mass media research: An introduction* (7th ed.). Belmont, CA: Wadsworth/Thomson.

Zhao, X. (2012). A reliability index (A_i) that assumes honest coders and variable randomness. Paper presented at the annual convention, Association for Education in Journalism and Mass Communication, Chicago.

Zhao, X., Liu, J. S., & Deng, K. (2012). Assumptions behind intercoder reliability indices. In C.T. Salmon (Ed.), *Communication yearbook 36* (pp. 419-480). New York: Routledge.

Figure 1. Linear trend of articles reporting simple agreement by year and journal

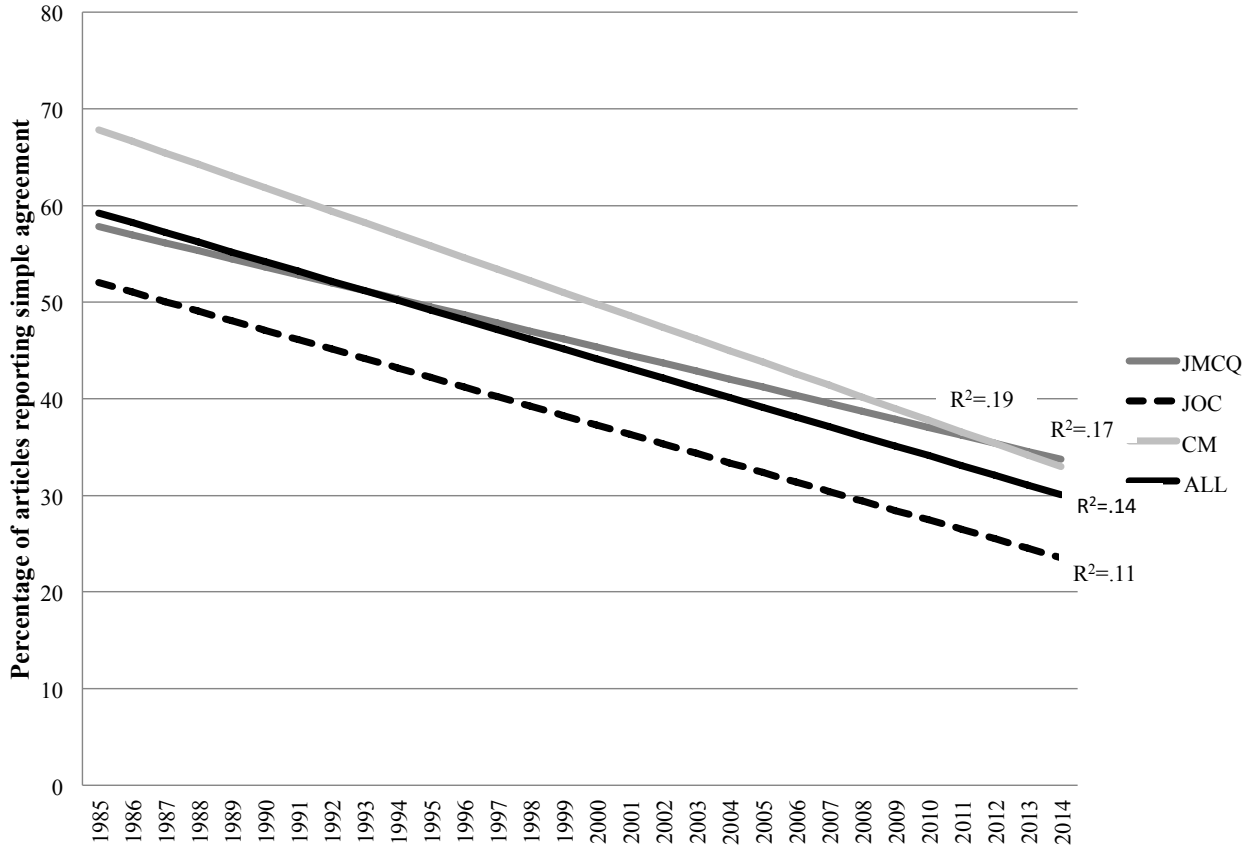

Figure 2. Linear trend of articles reporting reliability coefficient by year and journal

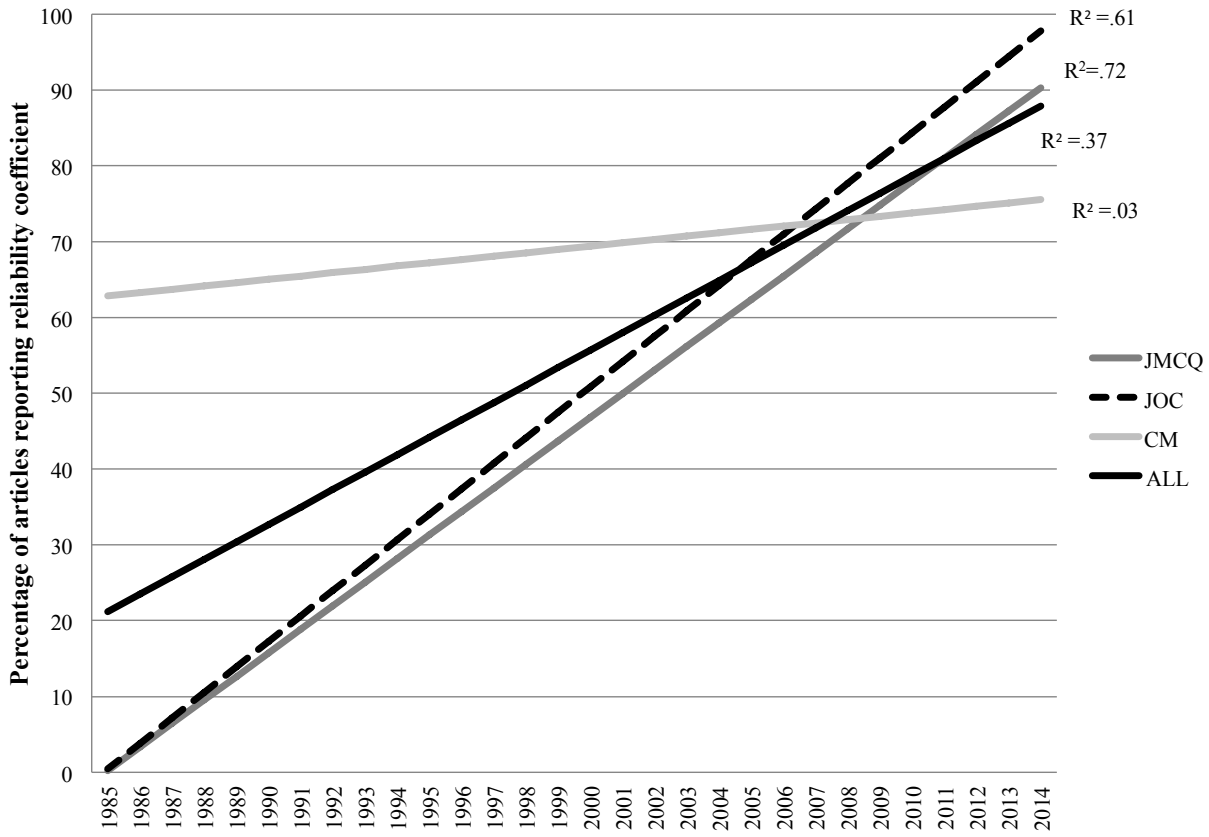

Figure 3. Linear trend of articles reporting reliability coefficient for every variable by year and journal

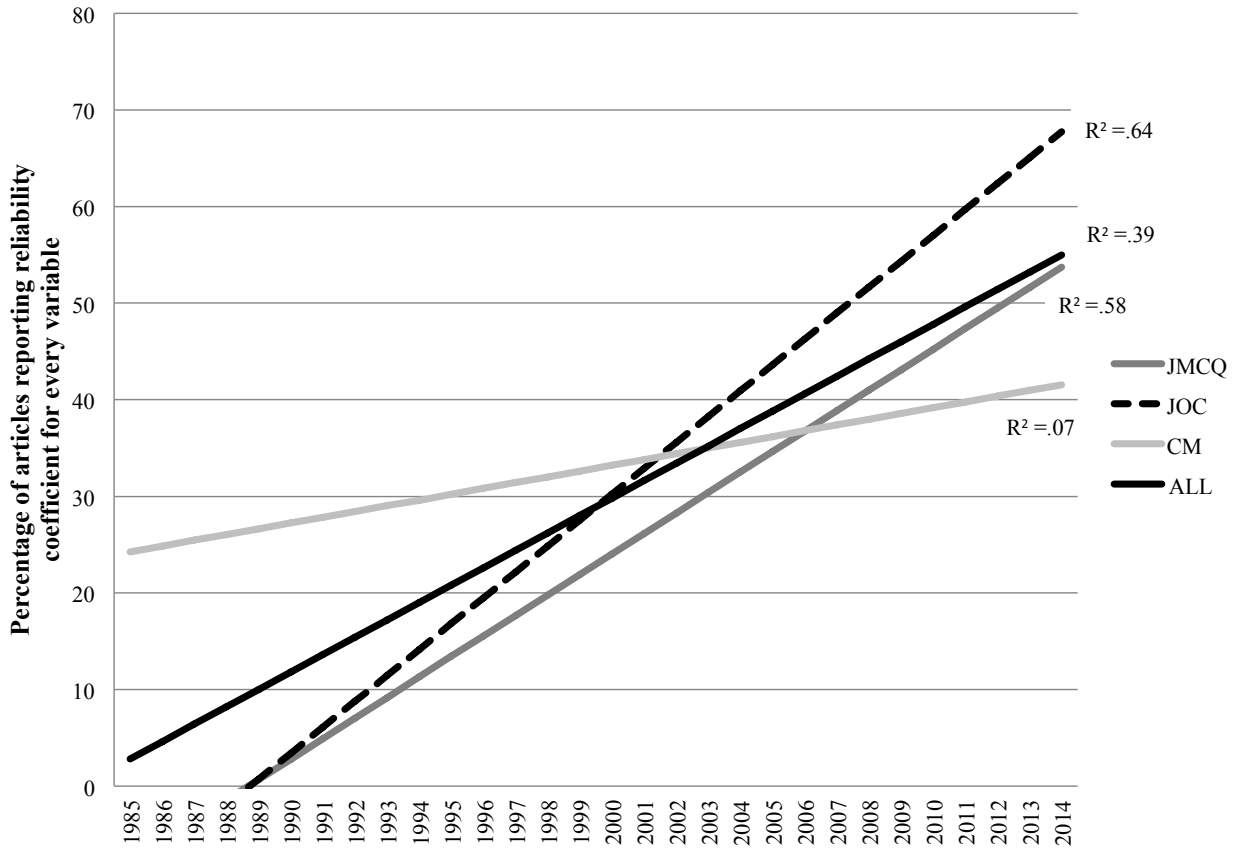

Figure 4. Linear trend of articles reporting reliability coefficients between .70 and .79 by year and journal

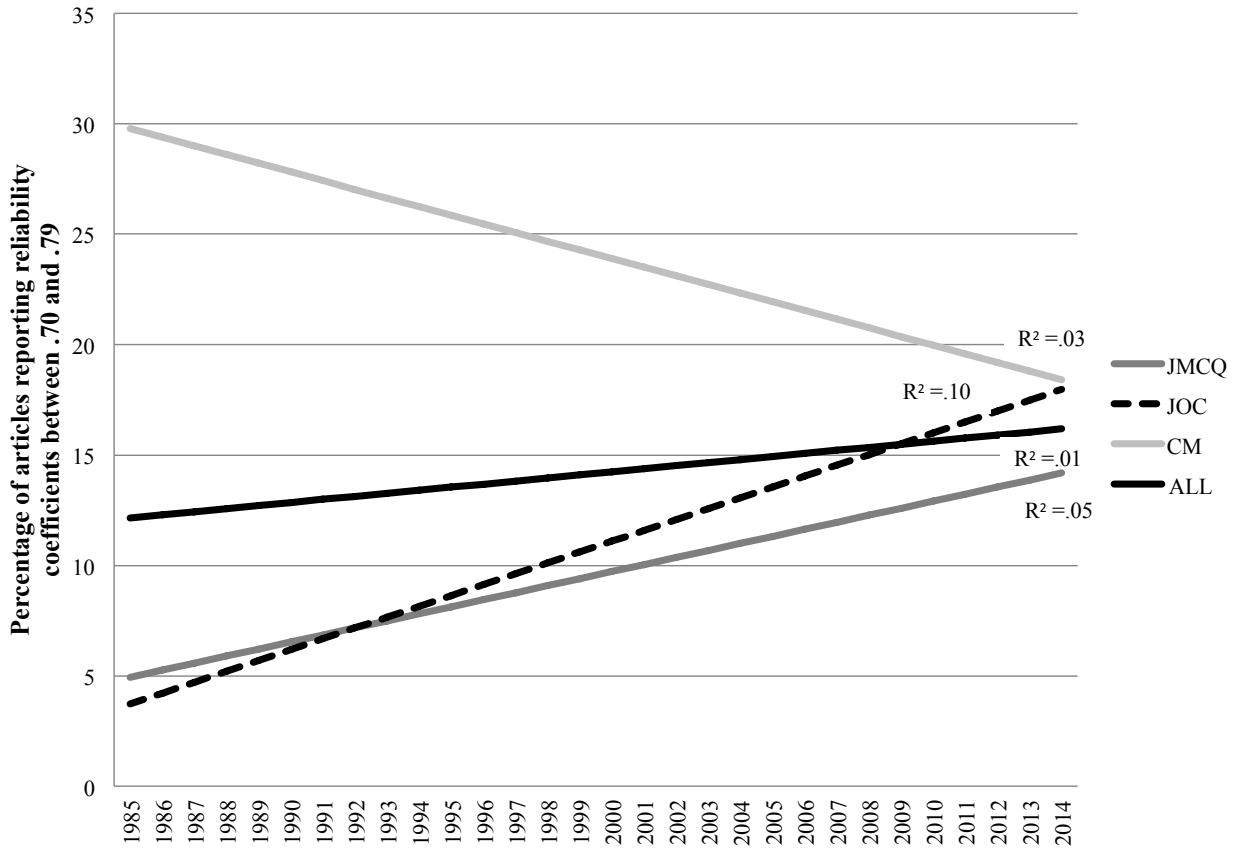

Figure 5. Linear trend of articles reporting reliability coefficients below .70 by year and journal

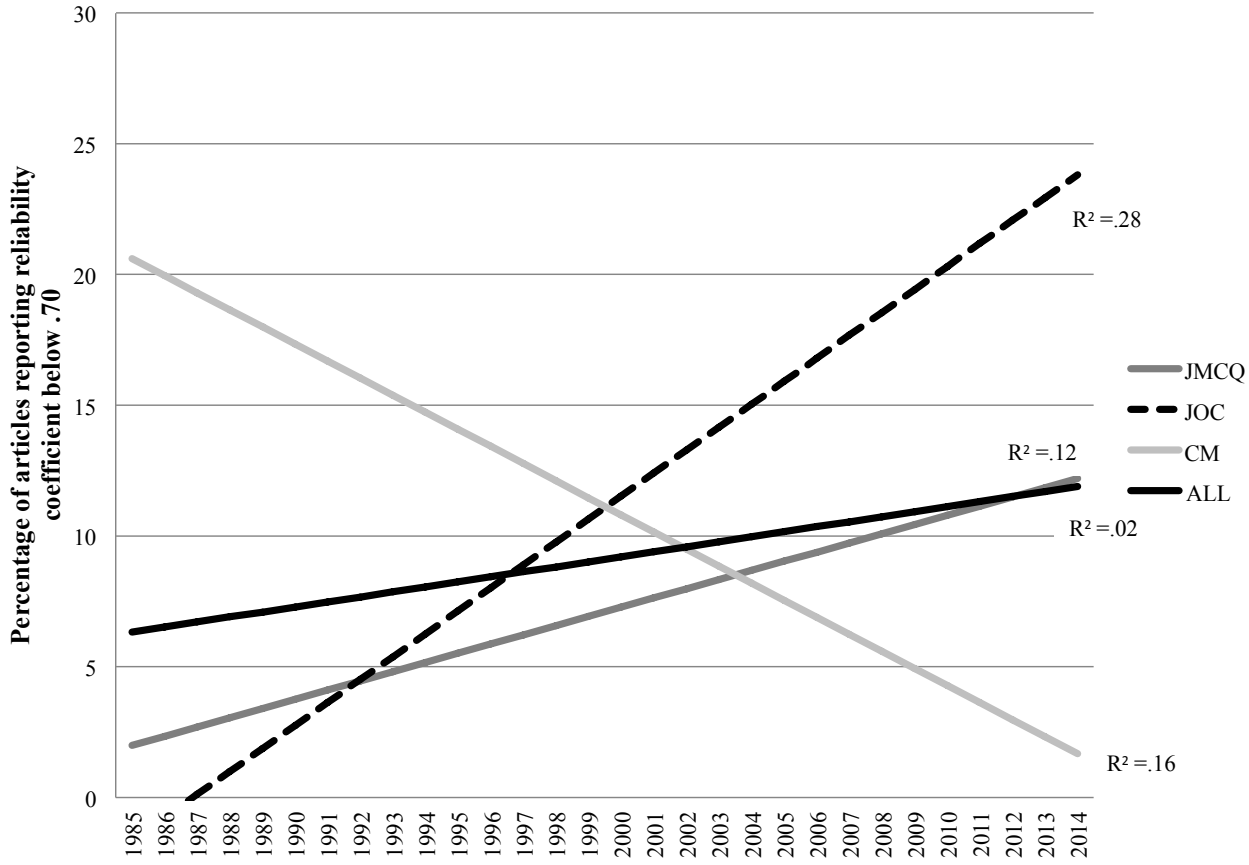

Table 1

Percentage of Articles Reporting Reliability by 5-year Strata and Journal

	Simple Agreement			Reliability Coefficients and Correlations			Chance-Corrected Reliability Coefficient			Reliability Coefficient for Every Variable			Reliability Coefficient Below .70			Reliability Coefficient Between .70 and .79		
	JMCQ	JoC	CM	JMCQ	JoC	CM	JMCQ	JoC	CM	JMCQ	JoC	CM	JMCQ	JoC	CM	JMCQ	JoC	CM
1985-1989	42	53	69	10	16	65	8	9	50	2	3	27	1	3	23	3	6	19
1990-1994	54	50	62	22	5	73	19	5	62	5	0	23	3	0	8	3	0	42
1995-1999	62	38	48	28	38	67	26	29	48	6	24	38	2	10	24	6	5	19
2000-2004	58	58	29	58	49	79	58	49	57	31	30	43	14	9	14	8	3	14
2005-2009	40	21	54	79	94	75	73	91	63	36	64	46	12	30	4	21	24	33
2010-2014	19	26	32	75	78	68	75	77	57	50	59	32	3	12	4	6	14	7

Table 2

Percentage of Articles within 5-year Strata Containing Specific Reliability Coefficients (across all journals); data presented as %(n)

Year	No Reliability Coefficient Reported	Correlation	Scott's Pi	Krippendorff's Alpha	Cohen's Kappa	Other Chance-Corrected Reliability Coefficient	Any Chance-Corrected Reliability Coefficient
1985-1989 (n = 174)	81.0 (141)	4.6 (8)	6.9 (12)	3.4 (6)	2.3 (4)	1.7 (3)	14.4 (25)
1990-1994 (n = 111)	69.4 (77)	4.5 (5)	13.5 (15)	0.9 (1)	9.0 (10)	2.7 (3)	26.1 (29)
1995-1999 (n = 89)	60.7 (54)	7.9 (7)	11.2 (10)	3.4 (3)	12.4 (11)	4.5 (4)	31.5 (28)
2000-2004 (n = 97)	39.2 (38)	6.2 (6)	22.7 (22)	9.3 (9)	14.4 (14)	8.2 (8)	54.6 (53)
2005-2009 (n = 90)	16.7 (15)	6.7 (6)	28.9 (26)	11.1 (10)	31.1 (28)	5.6 (5)	76.7 (69)
2010-2014 (n = 111)	23.4 (26)	5.4 (6)	16.2 (18)	28.8 (32)	25.2 (28)	2.7 (3)	73.0 (81)

Table 3

Part Correlations and Unstandardized Regression Coefficients for Independent Variables Associated with Percent of Articles with a Given Reliability Characteristic

Independent Variable	% Articles with Reliability Coefficient	% Articles with Coefficient for Each Variable	% Articles with Coefficients Below .70	% of Articles with Coefficients Between .70 and .79
Year	.606 (2.301)[c]	.624 (1.800)[c]	.130 (.192)	.072 (.139)
JMCQ Dummy	-.295 (-23.717)[c]	-.192 (-11.763)[a]	-.171 (-5.363)	-.354 (-14.463)[b]
JoC Dummy	-.253 (-20.323)[b]	-.079 (-4.833)	-.042 (-1.313)	-.298 (-12.170)[b]
R-Squared	.469[c]	.427[c]	.049	.150[b]

Note: N = 90; the unstandardized regression coefficient is in parenthesis.

[a] $p < .05$
[b] $p < .01$
[c] $p < .001$